




STORAGE CONTROL DEVICE

Patent number: JP11085408
Publication date: 1999-03-30
Inventor: YAMAMOTO AKIRA; NAKAMURA KATSUNORI; KIJIRO SHIGERU
Applicant: HITACHI LTD
Classification:
 - international: G06F3/06; G06F3/06
 - european:
Application number: JP19970248177 19970912
Priority number(s): JP19970248177 19970912

Also published as:

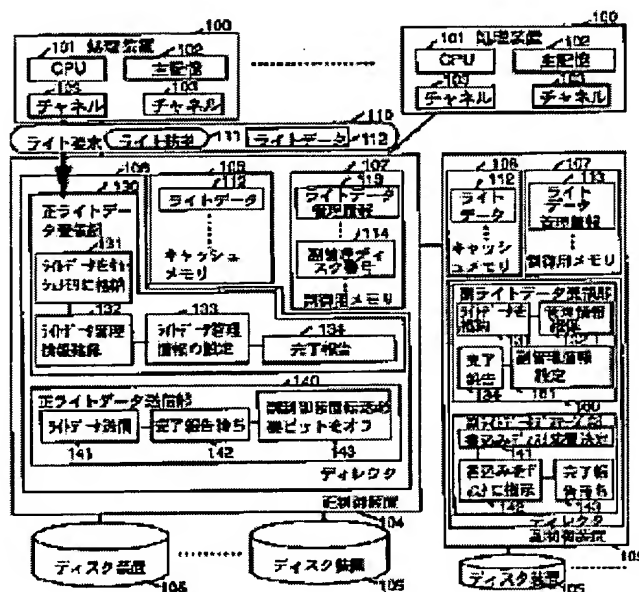
 E P0902370 (A2)
 US 6408370 (B2)
 US 2001029570 (A1)

Best Available Copy

Report a data error here

Abstract of JP11085408

PROBLEM TO BE SOLVED: To provide a function capable of minimizing the number of data transfers at the time of executing double writing between remote control devices, suppressing the deterioration of performance to the minimum even when the distance between the control devices is enlarged and preventing the halfway result of a transaction from being left, also to make it unnecessary to execute disk I/O processing for control information, and to attain high performance. **SOLUTION:** After returning write data completion report, a master control device 104 sends the reports directly to a sub-control device 109. The device 109 stores the received write data in a non-volatile memory to guarantee the data. Then, a time to be a certain reference is set up, all write data before the time are guaranteed and all write data after the time are discarded.



Data supplied from the esp@cenet database - Worldwide

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-85408

(43) 公開日 平成11年(1999) 3月30日

(51) Int.Cl.⁶
G 0 6 F 3/06識別記号
3 0 1
3 0 4F I
C 0 6 F 3/063 0 1 X
3 0 4 E

審査請求 未請求 請求項の数 8 O L (全 14 頁)

(21) 出願番号 特願平9-248177

(22) 出願日 平成9年(1997) 9月12日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 山本 彰

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72) 発明者 中村 勝彦

神奈川県小田原市国府津2880番地 株式会社日立製作所ストレージシステム事業部内

(72) 発明者 木城 茂

神奈川県小田原市国府津2880番地 株式会社日立製作所ストレージシステム事業部内

(74) 代理人 弁理士 小川 勝男

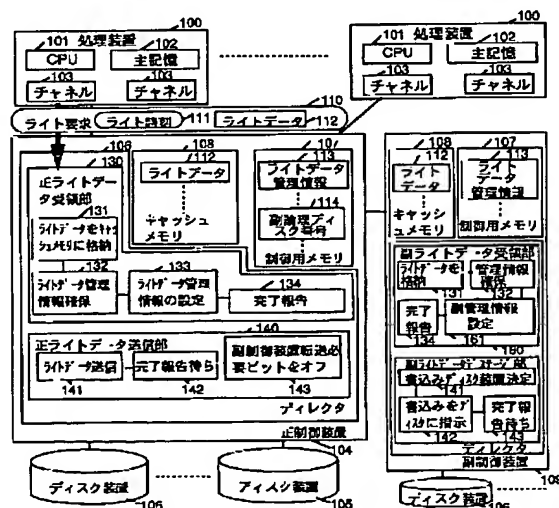
(54) 【発明の名称】 記憶制御装置

(57) 【要約】

【課題】本発明の課題は、遠隔地の制御装置間で2重書きを行う際、データ転送数を必要最小限に留め、制御装置間の距離が拡大しても、性能の劣化を微小に抑さえ、さらに、トランザクションの途中結果を残さないような機能を提供することである。加えて、制御情報のディスク入出力処理の実行も不必要とし、高性能化を図る点にある。

【解決手段】本発明では、正制御装置は、ライトデータを完了報告を返した後、直接副制御装置に送る。さらに、副制御装置は、受け取ったライトデータを不揮発メモリに格納することで、データ保証を行う。さらに、ある基準となる時刻を設け、この時刻以前のすべてのライトデータを保証し、この時刻より後のライトデータはすべて破棄できるようにする。

図 1



(2)

特開平11-85408

【特許請求の範囲】

【請求項1】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、前記複数の制御装置の内の1台以上の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、

前記処理装置に、前記ライト要求の完了報告を行う手段と、

前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送る手段とを有し、前記複数の制御装置の内の1台以上の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納する手段とを有し、前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。

【請求項2】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、前記複数の制御装置の内の1台以上の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、

前記処理装置に、前記ライト要求の完了報告を行う手段と、

前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送る手段とを有し、前記複数の制御装置の内の1台以上の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納する手段とを有し、前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有し、前記ライト時刻を参照して、前記ライトデータを前記不揮発メモリから破棄する手段を有することを特徴とする複合記憶装置システム。

【請求項3】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する2台の制御装置から構成される複合記憶装置システムであって、前記2台の制御装置の内の1台の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、前記処理装置に、前記ライト要求の完了報告を行う手段

と、

前記複合記憶システムを構成する他の制御装置に、前記ライト時刻順に前記ライトデータを送る手段とを有し、前記2台の制御装置の内の1台の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータを受け取り、前記不揮発メモリに格納する手段とを有し、

前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。

【請求項4】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、前記複数の制御装置の内の1台以上の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、

前記処理装置に、前記ライト要求の完了報告を行う手段と、

前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段とを有し、

前記複数の制御装置の内の1台以上の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、

前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、前記受領報告ライト時刻を受け取る手段とを有し、

前記ライト時刻と前記受領報告ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。

【請求項5】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、前記複数の制御装置の内の1台以上の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、

前記処理装置に、前記ライト要求の完了報告を行う手段と、

前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装

(3)

特開平11-85408

置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段を有し、

前記複数の制御装置の内の1台以上の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、

前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、受領報告ライト時刻を受け取る手段を有し、

前記ライト時刻をと前記受領報告ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段と、

前記ライト時刻と前記受領報告ライト時刻を参照して、前記不揮発性メモリから、前記ライトデータを消去する手段を有することを特徴とする複合記憶装置システム。

【請求項6】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、

前記複数の制御装置の内の1台以上の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、

前記処理装置に、前記ライト要求の完了報告を行う手段と、

前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送る手段とを有し、

前記複数の制御装置の内の1台以上の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納する手段とを有し、

受け取った前記ライト時刻に関する情報を他の制御装置に送る手段と、

前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。

【請求項7】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、

前記複数の制御装置の内の1台以上の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、

前記処理装置に、前記ライト要求の完了報告を行う手段

と、

前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段を有し、

前記複数の制御装置の内の1台以上の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、

前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、前記受領報告ライト時刻を受け取る手段を有し、

受け取った前記受領報告ライト時刻に関する情報を、他の制御装置に送る手段と、

前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。

【請求項8】それぞれの制御装置が1台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、

前記複数の制御装置の内の1台以上の制御装置が、処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、

前記処理装置に、前記ライト要求の完了報告を行う手段と、

前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段を有し、

前記複数の制御装置の内の1台以上の制御装置が、前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、

前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、受領報告ライト時刻を受け取る手段を有し、

受け取った前記受領報告ライト時刻に関する情報を、他の制御装置に送る手段と、

(4)

特開平11-85408

前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段と、
前記ライト時刻を参照して、前記不揮発性メモリから、前記ライトデータを消去する手段を有することを特徴とする複合記憶装置システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、異なった制御装置の間で、データを2重書きする機能に関する。特に、制御装置間の距離が長く、制御装置間のデータ転送に遅延が発生するような場合、本発明は、有効である。

【0002】

【従来の技術】本発明に関する公知例として、以下の技術が開示されている。

【0003】European Patent Application publication number 0671686A1は、では、遠隔地にある制御装置間のディスクの2重書きを行う技術が、開示されている。本発明では、一方の制御装置が、地震などの天災等により破壊されても、もう一方の制御装置のディスクでデータ保証が可能というものである。European Patent Application publication number 0671686A1では、ホスト計算機から、直接、ライトデータを受領する正側の制御装置は、遠隔地にある副側の制御装置へ、受領したライトデータを転送した後、ライトデータの受領完了を、ホスト計算機に報告する。この方法であると、正側と副側で完全にデータが一致するため、データ保証の点からは、非常に良い方法であった。しかし、制御装置間の距離拡大により、制御装置間のデータ転送時間は、非常に大きくなるため、遠距離時に、性能上の課題があった。

【0004】European Patent Application publication number 0672985A1でも、遠隔地にある制御装置間のディスクの2重書きを行う技術が、開示されている。European Patent Application publication number 0672985A1では、正側のホスト計算機から、直接、ライトデータを受領する正側の制御装置は、ライトデータ受領後直ちに、ライトデータの受領完了を、正側のホスト計算機に報告する。European Patent Application publication number 0672985A1では、さらに、正側の制御装置が受領したライトデータのコピーが、一度正側のホスト計算機に読みだされる。本発明では、当初正側のホスト計算機から受領するライトデータには、時刻が付与されている。時刻は、このライトデータを書き込むライト要求が発行された時刻を意味する。ライトデータのコピーが、正側のホスト計算機に読みだ

される時、ライト時刻も正側のホスト計算機に渡される。この後、正側のホスト計算機はライトデータのコピーとライト時刻を、副側のホスト計算機に送る。

【0005】ライトデータとライト時刻を受け取った副側のホスト計算機は、ライト時刻等の情報を、制御用のディスクに書き込む。さらに、各ライトデータに付与された時刻を参照し、ライト時刻順に、ライトデータを副側のディスクに書き込みを行う。

【0006】European Patent Application publication number 0672985A1で副側のホスト計算機が、上記のような処理を行う目的は、オンラインシステム等で標準的に使用されるトランザクションの途中結果を残さないようにするためである。例えば、口座Aから口座Bに預金を移すトランザクションを実行する場合、口座Aから預金を引き落とししたにもかかわらず、口座Bに預金を積み立てない状態を残さないようにすることが、トランザクションの途中結果を残さないということを意味する。通常、オンラインシステムでは、回復の単位は、トランザクションであるため、トランザクションの途中結果を残すことは、極めて重要な障害である。

【0007】次に、上記のような処理を実行すると、トランザクションの途中結果を残さないようにすることができることを簡単に説明する。2重書きを行っているディスクの中には、口座情報等のデータベースを格納したディスクと、トランザクションの更新履歴を残したジャーナルを格納したディスクがある。ホスト計算機がダウンすると、回復処理プログラムにより、ジャーナルが解析され、終了していないトランザクションの更新結果は、実行開始前の状態に戻される等の処理が、実行され、トランザクションの途中結果を残さないようにすることができる。副側の制御装置のディスクに書き込んだライトデータが、有効となるのは、最新のライトデータを格納した正側の制御装置が破壊してしまったような場合である。副側の制御装置には、最新のライトデータは格納されていないが、ある時刻までのライトデータは保証されていることになる。したがって、見かけ上、ホスト計算機が、ライトデータを保証している時刻に、ダウンしたのと等価な状態を作り出していることになる。したがって、副制御装置側のジャーナルを格納したディスクと、データベースを格納したディスクを用いて、ホスト計算機がダウンした時に実行される回復処理と同様の処理を実行することにより、トランザクションの途中結果を残さないようにすることができる。

【0008】特開平4-245342は、ディスク制御装置が不揮発性のキャッシュメモリを持ち、ライトアプタを行う、すなわち、ホスト計算機から受領したライトデータを不揮発性のキャッシュメモリに書き込み、完了報告を行う技術が開示されている。不揮発性のキャッシュメモリは、信頼性が高いため、ここにライトデータを

(5)

特開平11-85408

格納すれば、十分データ保証が可能となると判断できるためである。

【0009】

【発明が解決しようとする課題】European Patent Application publication number 0672985A1では、制御装置間の距離が拡大しても、若干のデータは、失われるものの、性能の劣化は、少ない。しかも、トランザクションの途中結果を残さない。

【0010】しかし、正側のホスト計算機がデータを読みだし、副側のホスト計算機にデータを転送するため、European Patent Application publication number 0671686A1のように、直接制御装置間で、ライトデータを受け渡す場合に比較し、データ転送が一度余分に実行される。さらに、MTなどの記憶媒体への入出力処理の実行も必要となる。

【0011】本発明の目的は、European Patent Application publication number 0672985A1のように、直接制御装置間で、ライトデータを受け渡し、しかも、制御装置間の距離が拡大しても、性能の劣化を微小に押さえ、しかも、トランザクションの途中結果を残さないような機能を提供することである。さらに、制御情報などのディスクへの入出力処理の実行も不必要とし、高性能化を図る。

【0012】

【課題を解決するための手段】以下、本発明が、以上述べてきた目的をいかに実現するかについて述べる。

【0013】本発明では、ホスト計算機は、正制御装置に、ライト要求を発行する際に、ライトデータにライト時刻を付与する。正制御装置は、ライトデータをホスト計算機から受け取ると、完了を報告する。この後、正制御装置は、副制御装置に、ライトデータとライト時刻を送る。この時、正の制御装置は、ライト時刻順に、ライトデータを、副の制御装置に送る。以上により、制御装置間の距離が拡大しても、性能の劣化を微小に押さえることができる。

【0014】副側の制御装置では、正側の制御装置から受け取ったライトデータとを、不揮発のキャッシュメモリに格納する。これにより、制御情報などのディスクへの入出力処理なしに、ライトデータのデータ保証が可能となる。

【0015】副側の制御装置では、受け取ったライト時刻を参照して、ある時刻までのライトデータを保証するようにする。これにより、トランザクションの途中結果を残さないようにすることが可能である。

【0016】

【発明の実施の形態】以下、本発明の実施例を説明する。まず、第1の実施例について説明する。

【0017】図1は、第1の実施例の概要を表す。第1の実施例における構成は、1台以上の処理装置100、1台の正制御装置104、正制御装置104に接続された1台以上のディスク装置105、1台の副制御装置109、副制御装置109に接続された1台以上のディスク装置105より構成する。処理装置100は、CPU101、主記憶102、チャネル103から構成される場合もある。正制御装置104は、制御用メモリ107、キャッシュメモリ108を含む。制御用メモリ107、キャッシュメモリ108は、不揮発化されているものとする。また、さらなる高信頼化のために、それぞれが2重化されていてもよい。キャッシュメモリ108、制御用メモリ107は、半導体メモリで構成されており、ディスク装置105に比べ、1桁から2桁高速なアクセスが可能である。正制御装置104は、処理装置100とディスク装置105の間のデータ転送を行う。さらに、本発明においては、正制御装置104は、副制御装置109の間のデータ転送を行う機能をもつ。あるいは、正制御装置104が1つ以上のディレクタ106を含み、各ディレクタ106が、処理装置200とディスク装置205との転送、副制御装置109との間のデータ転送を行ってもよい。また、副制御装置109の内部構成は、正制御装置104と同様である。

【0018】制御用メモリ107には、ライトデータ112に対応したライトデータ管理情報113が、作成される。

【0019】処理装置100は、正制御装置104にライト要求110を発行する時、ライトデータ112に、ライト時刻111に付与する。ライト時刻111は、本ライト要求110が発行された時刻を表しており、ライト時刻111により、処理装置100が発行したライト要求110順序を認識することができる。処理装置100が複数存在する場合、ライト時刻111は、処理装置100間で、共通のクロックなどを用い、異なった処理装置100で発行されたライト要求110の順序も、認識できるようになっているものとする。

【0020】図2は、ライトデータ管理情報113の構成である。ここでは、特に、本発明に直接関係する情報について説明する。なお、本発明では、処理装置100がライト要求110を発行する際、指定するディスクを論理ディスクとよぶ。論理ディスクID120は、対応するライトデータを書き込むよう、処理装置100から指示された論理ディスクの番号であり、ライト要求110に含まれる情報である。本発明では、処理装置100が認識している論理ディスクとディスク装置105（物理ディスク）は、1対1に対応している必要はない。図3に示すように、論理ディスクが、複数のディスク装置105上に定義されてもよい。また、論理ディスクに、冗長データを含ませ、RAID (Redundant Array of Inexpensive Dis

(6)

特開平11-85408

ks)構成にしてもよい。ライトアドレス121は、対応するライトデータを書き込む論理ディスク内のアドレスを示す情報(例えば、論理ディスクの先頭から1MByteの領域というような情報)で、ライト要求110に含まれる情報である。ライトデータ長122は、対応するライトデータの長さを表す情報であり、ライト要求110に含まれる情報である。以上の情報は、いずれも、通常のライト要求110に含まれる情報である。ライトデータポインタ123は、対応するライトデータ112へのポインタである。ライト時刻111については、すでに、説明したとおりである。ライト要求110に、ライト時刻111を付与することが本発明の特徴の1つである。副制御装置転送必要ビット124は、副制御装置109に対応するライトデータ112の転送が必要であることを表す情報である。

【0021】制御用メモリ108に含まれるもう1つの情報は、副論理ディスク番号114である。本情報は、正制御装置104の論理ディスク対応に存在する情報で、対応する論理ディスクの2重書きベアになっている副論理ディスクの番号、すなわち、副論理ディスクを格納している副制御装置109の番号と、副論理ディスクの副制御装置109内の論理ディスク番号を含む。もちろん、2重書きベアをもたない論理ディスクには、ヌル値が入るものとする。

【0022】副制御装置109の制御用メモリ109にも、ライトデータ管理情報113が含まれる。

【0023】フォーマットは、正制御装置104内のライトデータ管理情報113と同じでよい。ただし、副制御装置転送必要ビット124は、常にオフとなっているものとする。さらに、正論理ディスク番号131である。本情報は、副制御装置104の論理ディスク対応に存在する情報で、対応する論理ディスクの2重書きベアになっている正論理ディスクの番号、すなわち、正論理ディスクを格納している正制御装置104の番号と、正論理ディスクの正制御装置104内の論理ディスク番号を含む。もちろん、2重書きベアをもたない論理ディスクには、ヌル値が入るものとする。

【0024】正制御装置104の正ライトデータ受領部130は、処理装置100から、ライト要求110を受け取ったとき、動作を開始する。まず、受け取ったライトデータ112を、キャッシュメモリ108に格納する。(ステップ131)次に、正ライトデータ受領部140は、制御用メモリ108内のライトデータ管理情報113を、当該ライト要求対応に確保する。(ステップ132)さらに、ライト要求に含まれるライト時刻111等の情報を確保したライトデータ管理情報113に格納し、ライトデータポインタ123、副制御装置転送必要ビット124の設定を行う。(ステップ133)最後に、処理装置100に、ライト要求110の完了報告を行う。(ステップ134)以上の処理には、デ

ィスク装置105へのアクセスがないため、高速な応答が可能となる。ライトデータ112をディスク装置105に書き込む処理は、正制御装置104が後から実行する。この動作は、通常の制御装置の動作であるため、特に、詳細に記述しない。

【0025】正制御装置104の正ライトデータ送信部140は、ライトデータ112を副制御装置109に送る機能をもつ。まず、副制御装置転送必要ビット124が設定されているライトデータ管理情報113の中で、ライト時刻が最も以前であるライトデータ113を、対応する副論理ディスク番号130を参照して、2重書きベアが存在する副制御装置109へ送る。ライトデータ112の長さ、書き込みを行う副論理ディスク内のアドレスは、ライトデータ管理情報113内の情報を参照して指定する。(ステップ141)次に、副制御装置109からの完了報告をまつ。(ステップ142)完了報告が返ってくると、副制御装置転送必要ビット124をオフする。(ステップ143)この後、ステップ140へ戻り、次に送信すべきライトデータを見つける。

【0026】副制御装置109の副ライトデータ受領部160は、正制御装置104から、ライトデータ112を受け取った時動作する。副ライトデータ受領部160の処理内容は、ライトデータ管理情報113の設定において、副制御装置転送必要ビット124の設定を行わない(ステップ161)こと以外は、正ライトデータ受領部140の処理内容と同様である。

【0027】副制御装置109の副ライトデータデステージ部140は、ライトデータ112をディスク装置105に書き込む機能をもつ。まず、ライトデータ管理情報113の中で、ライト時刻が最も以前である順にいくつかのライトデータ113を、ディスク装置105に書き込むことを決定し、しかるべき計算を行い、書き込みを行うディスク装置105と書き込みアドレスを決める。この計算方法は、通常のRAID等で用いられる方法であるため、詳細には記述しない。(ステップ171)次に、ライトデータ112をディスク装置105に書き込むよう要求を複数並行して、ディスク装置105に発行する。(ステップ172)さらに、次に、ディスク装置105からの完了報告をまつ。(ステップ173)すべての要求の完了報告を受け取った後、ステップ170へ戻り、次にディスク装置105にデステージすべきライトデータ113を見つける。

【0028】正制御装置104から副制御装置109へのライトデータ113の送信順序が、ライト時刻111の順番であるため、副制御装置109では、ある時刻を基準に、それ以前のライトデータ113はすべて保持でき、それ以降のライトデータ113はまったく保持しないという状態を作り出すことができる。これにより、正制御装置104が破壊されても、副制御装置109側で、トランザクションの等中結果を残さない回復処理が

(7)

特開平11-85408

可能となる。また、副制御装置109側で、ライトデータ113、ライト時刻112等の制御情報は、キャッシュメモリ107、制御用メモリ113などの不揮発性の半導体メモリに保持されるため、性能上のオーバーヘッドは小さい。

【0029】以上説明してきた内容は、正制御装置104から副制御装置109へのライトデータ112の転送がシリアルライズされているため、十分な性能が得られない可能性がある。図4は、正制御装置104から副制御装置109へのライトデータ112の転送を並列に実行した場合の動作を表している。各処理部で、転送がシリアルライズされている場合と変更があるのは、正ライトデータ送信部a300、正基準時刻送信部170、副基準時刻受信部180、副ライトデータデステージ部a310と正障害時データ破棄部190である。以下、正ライトデータ送信部a300の処理フローについて説明する。まず、副制御装置転送必要ビット124が設定されているライトデータ管理情報113の中で、ライト時刻が最も以前である順にいくつかのライトデータ113を、対応する副論理ディスク番号130を参照して、2重書きペアが存在する副制御装置109へそれぞれ並列に転送する。(ステップ301)次に、副制御装置109から、それぞれの完了報告が送られてくるのをまつ。(ステップ302)すべての完了報告が返ってくると、対応するライトデータ管理情報113の中の副制御装置転送必要ビット124をオフする。(ステップ303)この後、ステップ150へ戻り、次に送信すべきライトデータ112を見つける。

【0030】ライトデータ112の転送を並列に実行すると、副制御装置109側で保持されるライトデータ112のライト時刻111の順序がくるう可能性がある。したがって、副制御装置109がデステージしてよいライトデータ112を決めるための判断基準となるライト時刻111を認識する必要がある。この場合、デステージしてよいライトデータ112は、正制御装置104の中で、副制御装置転送必要ビット124がオンになっているライトデータ管理情報113の中で、最も以前のライト時刻111を基準時刻として、この基準時刻より以前のライト時刻111をもつライトデータ112ということになる。というのは、この基準時刻より以前のライト時刻111をもつライトデータ112はすべて、副制御装置109側に保持されていることになるためである。一方、この基準時刻より後のライト時刻111をもつライトデータ112は、まだデステージしてはまじいライトデータ112であり、正制御装置104が破壊された場合、これらのライトデータ112はデステージせず、破棄する必要がある。

【0031】正基準時刻送信部170は、副制御装置109に上述したデステージしてよい基準時刻を送信する機能をもつ。基準時刻は、上述したように、副制御装置

転送必要ビット124がオンになっているライトデータ管理情報113の中で、最も以前のライト時刻111である。

【0032】副基準時刻受信部180は、正制御装置104から受信した基準時刻を、デステージ許可時刻185として、制御用メモリ108に格納する。

【0033】図4は、正制御装置104から副制御装置109へのライトデータ112の転送を並列に実行した場合の副ライトデータデステージ部a310の処理フローである。図1に示した処理フローとは異なるのは、デステージするライトデータ113を選択する条件に、デステージ許可時刻185より以前のライト時刻111であるかどうかという条件が入ることだけである(ステップ311)。

【0034】正障害時データ破棄部197は、正制御装置104が破壊された時、デステージ許可時刻185から後のライト時刻111をもつライトデータ112を破棄する機能をもつ(ステップ191)。

【0035】次に、第2の実施例について説明する。

【0036】図5は、第2の実施例の概要を表す。第2の実施例と第1の実施例の相違は、正制御装置104と副制御装置109の数である。第1の実施例では、正制御装置104の数が1台で、副制御装置109の数が1台であった。一方、第2の実施例では、正制御装置104の数が2台以上で、副制御装置209の数が1台である。

【0037】正制御装置104が複数存在すると、副制御装置109側で、それぞれの正制御装置104から受け取っているライトデータ112のライト時刻111にずれが生ずる。一方の正制御装置104(例えば、正制御装置a)から受け取った最も最近のライト時刻111、この時刻を時刻a、もう一方の正制御装置104(例えば、正制御装置b)から受け取った最も最近のライト時刻111、この時刻を時刻bとする。この場合、時刻aより、時刻bの方が、以前の時刻であるとする。正制御装置a側に、時刻aより最近で、時刻bより以前のライトデータ113を保持している可能性がある。すでにのべたように、トランザクションの途中結果を残さないようにするには、ある基準時刻以前のライト時刻112をもつライトデータ113はすべて保証し、基準時刻以降のライト時刻112をもつライトデータ113はすべて破棄する必要がある。したがって、時刻a以前のライト時刻111をもつライトデータ112が、副制御装置109でデステージしてよいライトデータ112ということになる。

【0038】以上に対応して、副制御装置109の制御用メモリ108には、正制御装置ライト許可時刻500がある。正制御装置ライト許可時刻500は正制御装置104ごとに存在する情報で、対応する正制御装置104から受け取った最も最近のライト時刻111が格納さ

(8)

特開平11-85408

れている。したがって、上述したように、これらの正制御装置ライト許可時刻500の中で、もっとも以前の時刻を基準時刻として、この基準時刻以前のライト時刻111をもったライトデータ112が、副制御装置109でデステージしてよいライトデータ112ということになる。

【0039】以下、本実施例でも、1台の正制御装置104から副制御装置109へのライトデータ112の転送を並列に実行した場合の各処理部の内容について述べる。もちろん、1台の正制御装置104から副制御装置109へのライトデータ112の転送をシリアルライズして実行した場合についても、本実施例は有効である。

【0040】正制御装置104の各処理部の処理フローは第1の実施例で、ライトデータ112の転送を並列に実行した場合（図3の処理）の処理フローと同様である。

【0041】次に、副制御装置109の各処理部の処理フローの説明を行う。

【0042】第2の実施例における副ライトデータデステージ部b510の処理フローについて説明する。ここでは、第2の実施例における副ライトデータデステージ部520の処理フローが、第1の実施例における副ライトデータデステージ部170の処理フローと異なる点について説明する。第2の実施例における副ライトデータ受領部510の処理内容は、デステージするライトデータ112を選択する際、対応するライト時刻111が、すべての正制御装置ライト許可時刻500より以前であるかをチェックして、条件を満たすライトデータ112を選択する点である。（ステップ511）これ以外は、第2の実施例における副ライトデータデステージ部b510の処理フローは、第1の実施例における副ライトデータデステージ部170の処理フローと同様である。

【0043】副基準時刻受信部b520は、正制御装置104から受信した基準時刻を、基準時刻を送信してきた正制御装置104に対応する正制御装置ライト許可時刻500に設定する。

【0044】本実施例においては、正障害時データ破棄部b530が、破棄するライトデータ112は、対応するライト時刻111が、すべての正制御装置ライト許可時刻500より以前であるという条件を満足しないライトデータ112である。（ステップ531）次に、第3の実施例について説明する。

【0045】図6は、第の実施例の概要を表す。第3の実施例と第2の実施例の相違も、正制御装置104と副制御装置109の数である。第3の実施例では、正制御装置104の数が2台以上で、副制御装置109の数は1台以上である。この場合、すべての正制御装置104と副制御装置109のペアがお互いに、接続されている必要はない。

【0046】副制御装置109が複数存在すると、トラ

ンザクションの途中結果を残さないようにするには、各副制御装置109間で、デステージするライトデータ112を選択する際に用いる基準時刻を共通にする必要がある。これは、データベースやジャーナルが複数の副制御装置109間に分散している可能性があるためである。

【0047】本実施例では、デステージするライトデータ112を選択する際に用いる基準時刻を決定する機能を、マスタ副制御装置700に持たせる。したがって、マスタ副制御装置700とそれ以外の副制御装置109の間は、データ転送路で接続されている。データ転送路が故障すると、各副制御装置109間で、デステージするライトデータ112を選択する際に用いる基準時刻を共通化することができなくなるため、データ転送路は多重化しておくことが望ましい。本実施例では、デステージするライトデータ112を選択する際に用いる基準時刻を決定する機能を、マスタ副制御装置700に持たせたが、基準時刻を決定する機能を、特定の副制御装置109に持たせず、各副制御装置109に分散させる方法（例えば、交代で、各副制御装置109が基準時刻を決定するような方法）をとっても、本発明は有効である。

【0048】以上に対応して、マスタ副制御装置700の制御用メモリ108には、副制御装置ライト時刻701がある。副制御装置ライト時刻701は、マスタ副制御装置700も含めた副制御装置109対応の情報である。各副制御装置ライト時刻701は、各副制御装置109から、マスタ副制御装置700が、適当な周期で、その副制御装置109内のすべての正制御装置ライト許可時刻500の中で、もっとも以前の時刻（実施例2で、副制御装置109がライトデータの選択の際、基準とした時刻）受け取る情報である。

【0049】マスタライト時刻702は、第3の実施例において、各副制御装置109がライトデータの選択の際、基準とする時刻である。マスタライト時刻702は、マスタ副制御装置700が、適当な周期で、すべての副制御装置ライト時刻701を参照して、もっとも以前の時刻を選択して、この時刻を設定する。選択した時刻以前のライト時刻111をもったすべてのライトデータ112は副制御装置109に保持されていることになる。このため、この条件を満足するライトデータ112を保証して、満足しないライトデータ112はすべて破棄することにより、トランザクションの途中結果を残さないようにすることができる。

【0050】以下、本実施例でも、1台の正制御装置104から副制御装置109へのライトデータ112の転送を並列に実行した場合の各処理部の内容について述べる。もちろん、1台の正制御装置104から副制御装置109へのライトデータ112の転送をシリアルライズして実行した場合についても、本実施例は有効である。

【0051】正制御装置104の各処理部の処理フロー

(9)

特開平11-85408

は第2の実施例とほとんど同様である。もちろん、正基準時刻送信部170はライトデータ112を送信する各副制御装置109に基準時刻を送る機能をもつ。ただし、送信する基準時刻は、送信を行う副制御装置109に対応するすべてのライトデータ112の副制御装置転送必要ビット124がオンになっているライトデータ管理情報113の中で、最も以前のライト時刻111である。

【0052】正制御装置104の各処理部の処理フローの中で、第2の実施例と異なるのは、副制御装置109に、副ライト時刻送信部710、副ライトデータデステージ部c720、正障害時データ破棄部cを含む点である。また、マスタ副制御装置700は、マスタ副ライト時刻受信部711、マスタライト時刻計算部712、マスタ副ライト時刻送信部713730を含む点である。

【0053】副ライト時刻送信部710は、適当な周期で、その副制御装置109内のすべての正制御装置ライト許可時刻500の中で、もっとも以前の時刻180を、マスタ副制御装置700のマスタ副ライト時刻受信部711に送る。マスタ副制御装置700以外の副制御装置109の副ライト時刻送信部710は、副制御装置109間のデータ転送路を利用する。マスタ副制御装置700の副ライト時刻送信部710は、マスタ副制御装置700の通信手段を利用する。

【0054】マスタ副ライト時刻受信部711は、副ライト時刻送信部710から受信した時刻を、当該時刻を送ってきた副制御装置109に対応する副制御装置ライト時刻701に、設定する。

【0055】マスタライト時刻計算部712は、適当な周期で、すべての副制御装置ライト時刻701を参照して、もっとも以前の時刻を選択して、この時刻を、マスタライト時刻702に設定する。

【0056】マスタ副ライト時刻送信部713は、各副制御装置109の副ライトデータデステージ部720、正障害時データ破棄部730からの要求にしたがって、適当な周期で、マスタライト時刻702に設定された時刻を送る。マスタ副制御装置700以外の副制御装置109には、副制御装置109間のデータ転送路を利用する。マスタ副制御装置700からの要求には、マスタ副ライト時刻送信部713は、マスタ副制御装置700の通信手段を利用する。

【0057】副ライトデータデステージ部720が、第2の実施例と異なる点は、デステージを行うライトデータ112を選択する際、マスタ副ライト時刻送信部713から基準となる時刻を受信し、この時刻より以前のライト時刻111をもつライトデータ112をデステージの対象として選択する点である（ステップ721）。

【0058】正障害時データ破棄部c730が、第2の実施例と異なる点は、キャッシュメモリ107から破棄するライトデータ112を選択する際、マスタ副ライト

時刻送信部713から基準となる時刻を受信し、この時刻より以前のライト時刻111をもつライトデータ112以外のライトデータ112を破棄の対象として選択する点である。（ステップ731）本実施例では、マスタ制御装置700が、副制御装置109から基準となる時刻を計算するのに必要な情報を受け取ったが、図7に示すように、正制御装置104から受け取るようにしてもよい。この場合、マスタ副制御装置700の制御用メモリ108には、マスタ正制御装置ライト時刻800がある。マスタ正制御装置ライト時刻800は、正制御装置104対応の情報である。マスタ正制御装置ライト時刻800は、各正制御装置104から、マスタ副制御装置700が、適当な周期で、その正制御装置109内で、オン状態の副制御装置転送必要ビット124を含むすべてのライトデータ管理情報113の中で、最も以前のライト時刻111を受け取り、設定を行う情報である。マスタライト時刻702には、マスタ副制御装置700が、適当な周期で、すべての正制御装置ライト時刻701を参照して、もっとも以前の時刻を選択して、この時刻を設定する。デステージ、データ破棄の際、基準となる時刻として用いる時刻が、マスタライト時刻702である点は、同様である。

【0059】

【発明の効果】本発明の目的は、遠隔地の制御装置間で2重書きを行う際、直接制御装置間で、ライトデータを受け渡すことにより、データ転送数を必要最小限に留め、しかも、制御装置間の距離が拡大しても、性能の劣化を微小に押さえ、さらに、トランザクションの途中結果を残さないような機能を提供することである。加えて、制御情報のディスク入出力処理の実行も不必要とし、高性能化を図る。

【図面の簡単な説明】

【図1】第1の実施例の概要図。

【図2】ライトデータ管理情報のフォーマット。

【図3】正制御装置から副制御装置へのライトデータの転送を並列に実行した場合の正ライトデータ送信部の処理フロー図。

【図4】正制御装置から副制御装置へのライトデータの転送を並列に実行した場合の副ライトデータデステージ部の処理フロー図。

【図5】第2の実施例の概要図。

【図6】第3の実施例の概要図。

【図7】第3の実施例において、マスタ副制御装置が、正制御装置からライト時刻に関する情報を収集した場合の処理概要図。

【符号の説明】

104…正制御装置、107…制御用メモリ、108…キャッシュメモリ、109…副制御装置、111…ライト時刻、112…ライトデータ、113…ライトデータ管理情報、124…副制御装置転送必要ビット、130

(10)

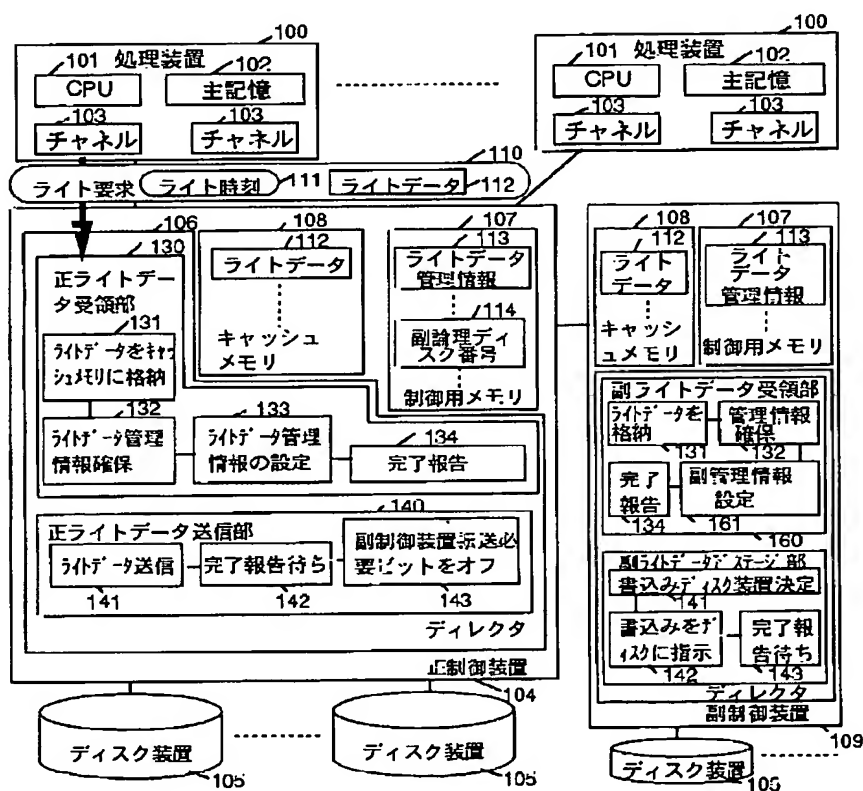
特開平11-85408

…正ライトデータ受領部、140…正ライトデータ送信部、150…副ライトデータ受領部、160…副ライトデータデステージ部、170…正基準時刻送信部、180…副基準時刻受信部、185…デステージ許可時刻、190…正障害時データ破棄部、300…正ライトデータ送信部a、310…副ライトデータデステージ部a、500…正制御装置ライト許可時刻、510…副ライトデータデステージ部b、520…副基準時刻受信部b、

530…正障害時データ破棄部b、700…マスタ副制御装置、701…副制御装置ライト時刻、702…マスタライト基準時刻、710…副ライト時刻送信部、711…マスタ副ライト時刻受信部、712…マスタ副ライト時刻計算部、713…マスタ副ライト時刻送信部、710…副ライトデータデステージ部c、720…正障害時データ破棄部c、800…正制御装置ライト時刻。

【図1】

図1



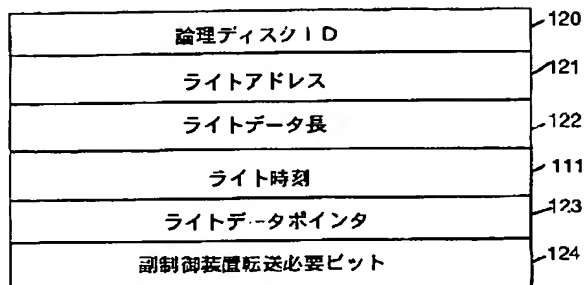
(11)

特開平11-85408

【図2】

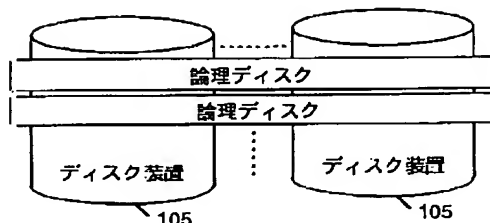
図 2

113



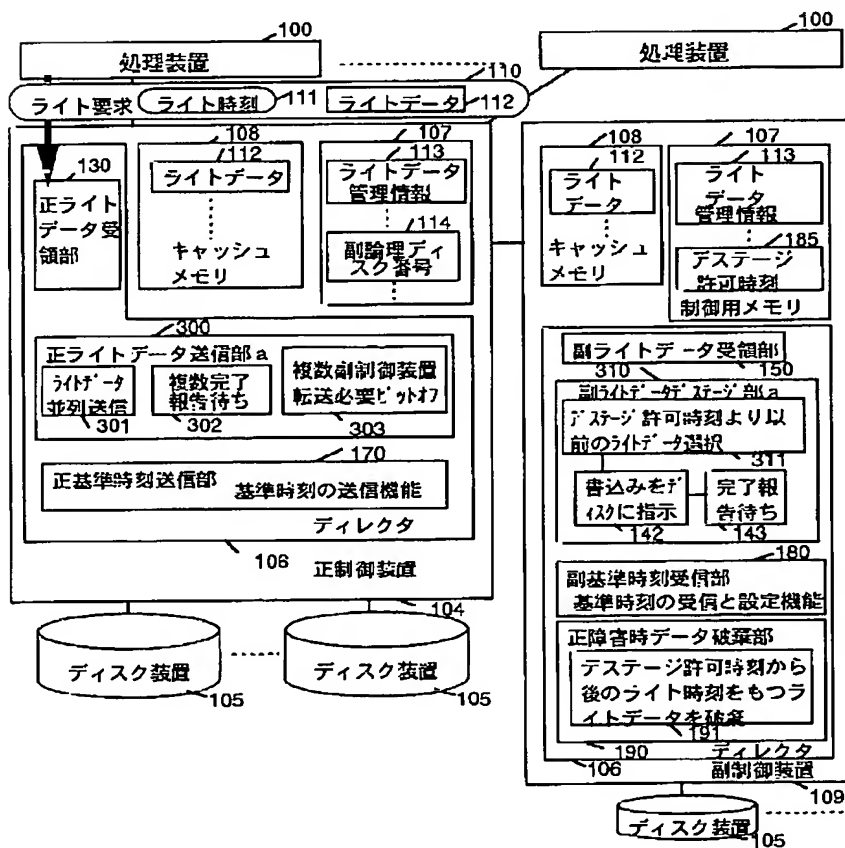
【図3】

図 3



【図4】

図 4

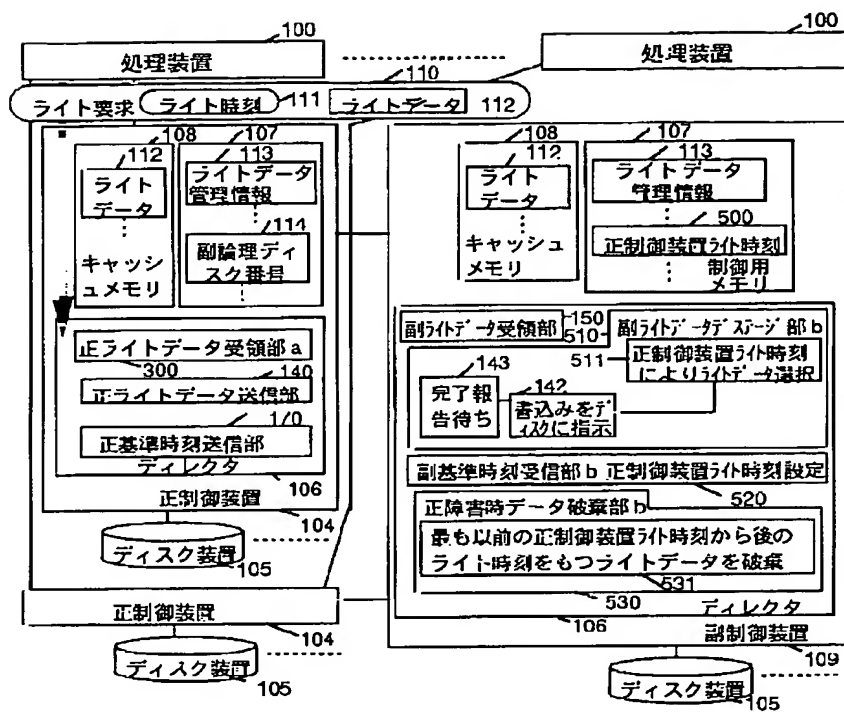


(12)

特開平11-85408

【図5】

図5

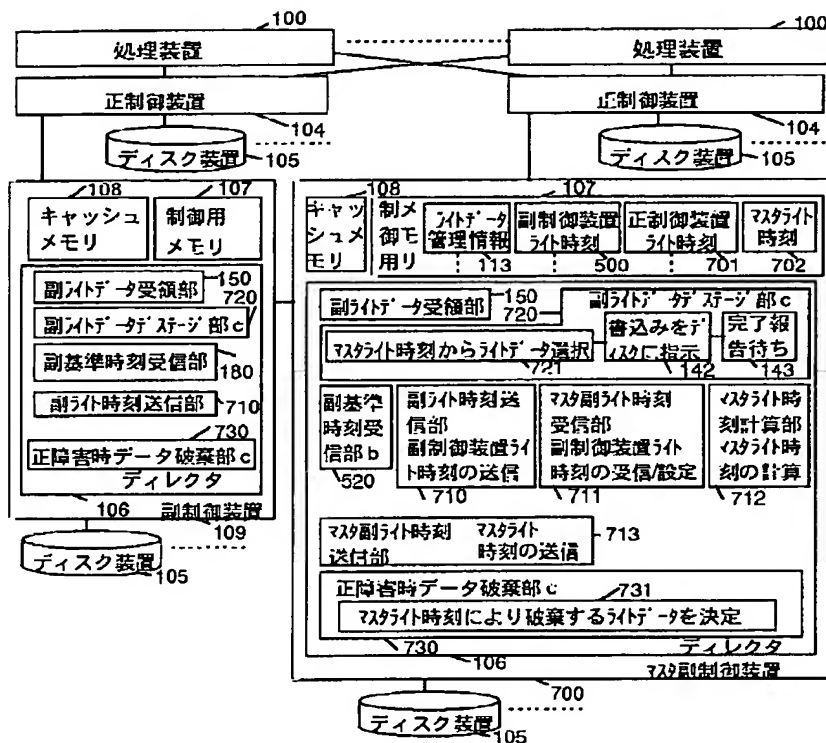


(13)

特開平11-85408

【図6】

図 6

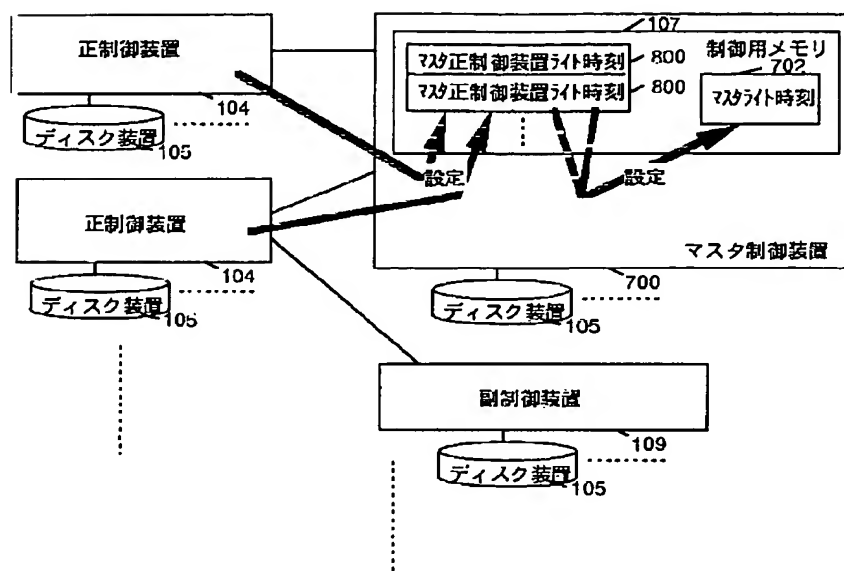


(14)

特開平11-85408

【図7】

図7



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.